Erasmus University Rotterdam

SPEED: A Semantics-Based Pipeline for Economic Event Detection

Introduction

Nowadays, emerging news on economic events such as acquisitions has a substantial impact on the financial markets. Therefore, it is important to be able to automatically and accurately **identify economic events in news items** in a timely manner. For this, one has to be able to process a large amount of heterogeneous sources of unstructured data in order to extract knowledge useful for guiding decision making processes. Because we hypothesize that domain-specific information captured in semantics facilitates detection of relevant concepts, we propose a **Semantics-based Pipeline for Economic Event Detection** (SPEED). In our approach, we aim to extract financial events from emerging news gathered from RSS feeds and to annotate these with machine-understandable meta-data, while retaining a speed that is high enough to make real-time use possible.

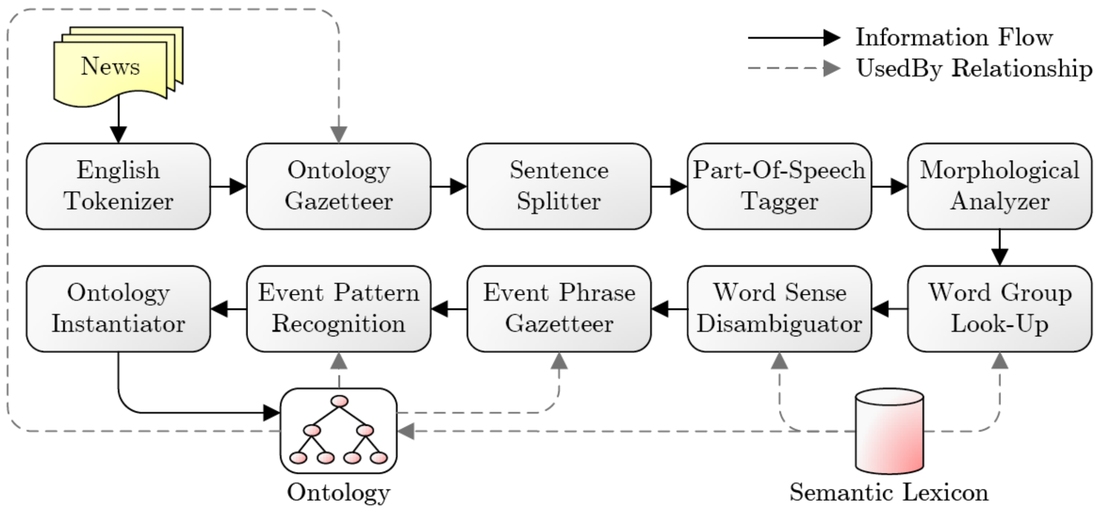

Fig. 1. SPEED design.

Economic Event Detection based on Semantics

Our framework is modeled as a **pipeline** and is driven by a financial **ontology** developed by domain experts, containing information on the NASDAQ-100 companies, extracted from Yahoo! Finance. Many concepts in this ontology stem from a semantic lexicon (e.g., WordNet), but another significant part of the ontology consists of concepts representing named entities (i.e., proper names).

Figure 1 depicts the architecture of the pipeline. In order to identify relevant concepts and their relations, the **English Tokenizer** is employed, which splits text into tokens (which can be for instance words or numbers) and subsequently applies linguistic rules in order to split or merge identified tokens. These tokens are linked to ontology concepts by means of the **Ontology Gazetteer**.

Subsequently, the **Sentence Splitter** groups the tokens in the text into sentences. These sentences are used for discovering the grammatical structures in a corpus by determining the type of each word token by means of the **Part-Of-Speech Tagger**. As words can have many forms that have a similar meaning, the **Morphological Analyzer** subsequently reduces the tagged words to their lemma as well as an affix.

A word can have multiple meanings and a meaning can be represented by multiple words. Hence, the framework needs to tackle word sense disambiguation tasks, given part-of-speech tags, lemmas, etcetera. To this end, first of all, the **Word Group Look-Up** component combines words into maximal word groups. Subsequently, the **Word Sense Disambiguator** determines the word sense of each word group by exploring the mutual relations between senses of word groups using semantic graphs. The senses are determined based on the number and type of detected semantic interconnections in a labeled directed graph representation of all senses of the considered word groups.

After disambiguating word group senses, the text can be interpreted by introducing semantics, which links word groups to an ontology, thus capturing their essence in a meaningful and machine-understandable way. Therefore, the **Event Phrase Gazetteer** scans the text for specific (financial) events, by utilizing a list of phrases or concepts that are likely to represent some part of a relevant event. Events thus identified are then supplied with available additional information by the **Event Pattern Recognition** component, which matches events to lexico-semantic patterns that are subsequently used for extracting additional information. Finally, the knowledge base is updated by inserting the identified events and their extracted associated information into the ontology by means of the **Ontology Instantiator**.

The performance of our proposed pipeline has been assessed by means of the **General Architecture for Text Engineering** (GATE). This general purpose framework for information extraction tasks provides the possibility to construct processing pipelines from components that perform specific tasks, e.g., linguistic, syntactic, and semantic analysis tasks. The **Java-implementation** of the proposed framework uses some default GATE components, such as the English Tokenizer, Sentence Splitter, Part-Of-Speech Tagger, and the Morphological Analyzer, which generally suit our needs. Furthermore, we extended the functionality of some other GATE components (e.g., ontology gazetteering), and also implemented additional components to tackle the disambiguation process.

Initial **results** on a test corpus of 200 news messages fetched from the Yahoo! Business and Technology RSS feeds show **fast** gazetteering of about 1 second and a **precision** and **recall** for concept identification in news items of **86%** and **81%**, respectively, which is comparable with existing systems. Precision and recall of **fully decorated events** result in lower values of approximately **62%** and **53%**, as they rely on multiple concepts that have to be identified correctly.

Conclusions

Our framework aims to extract financial events from news articles (announced through RSS feeds) and to annotate these with meta-data, while maintaining a speed that is high enough to enable real-time use. The main components of the framework introduce some novelties, as they are **semantically enabled**. Furthermore, pipeline outputs also make use of semantics, which introduces a potential **feedback** loop, making event identification a more adaptive process. Finally, we briefly touched upon the implementation of the framework and initial test results on the basis of emerging news. The established fast processing time and high precision and recall provide a good basis for future work. The **merit** of our pipeline is in the use of **semantics**, enabling not only a good performance, but also broader application interoperability.

Frederik Hogenboom, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, Otto van der Meer, Kim Schouten, and Damir Vandic
{fhogenboom, hogenboom, frasincar, kaymak}@ese.eur.nl,
{276933rm, 288054ks, 305415dv}@student.eur.nl

Econometric Institute
Erasmus School of Economics
Erasmus University Rotterdam
P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands
Phone: +31 (0)10 408 8907
Fax: +31 (0)10 408 9031
http://www.eur.nl/ese/english/

ERASMUS UNIVERSITEIT ROTTERDAM